

The exponential family in abstract information theory

Jan Naudts

Ben Anthonis

University of Antwerp, Physics Department
Universiteitsplein 1, 2610 Wilrijk-Antwerpen, Belgium
jan.naudts@ua.ac.be, ben.anthonis@ua.ac.be

February 22, 2013

Abstract

We introduce generalized notions of a divergence function and a Fisher information matrix. We propose to generalize the notion of an exponential family of models by reformulating it in terms of the Fisher information matrix. Our methods are those of information geometry. The context is general enough to include applications from outside statistics.

1 Introduction

The literature contains several generalizations of the concept of models belonging to the exponential family [1]. See for instance [2, 3, 4, 5, 6, 7]. The present work gives such a definition in a context of an abstract information theory, which is not necessarily based on probability. The main tools are those of information geometry [8], in particular generalized divergence functions [2, 3, 4, 9, 10, 11]. They can be used to define a generalized Fisher information matrix and generalized exponential families (Definitions 1 and 2 in Section 2).

The motivation for the present work comes from physics. Applications of the new definitions in the context of classical and of quantum mechanics will follow in a separate publication [12]. A preliminary write-up of the present work, including one non-statistical example, is found in [13].

The next section introduces a generalized divergence in an abstract setting. The Bregman divergence, discussed in Section 3, is an important subcase. Section 4 introduces our definitions of generalized Fisher information and of generalized exponential families. Sufficient conditions for a family to belong to a generalized family follow in Section 5. The final two sections show how our definitions relate to other generalizations found in the literature.

2 Definitions

The abstract information framework $\mathbb{X}, \mathbb{M}, \mathbb{Q}, \mu$ consists of a topological space \mathbb{X} , a differentiable manifold \mathbb{M} , and a linear space \mathbb{Q} of real functions of \mathbb{X} . In

addition there is given a continuous map $\mu : \mathbb{X} \rightarrow \mathbb{M}$. The space \mathbb{X} contains data sets. The map μ associates a model point with each data set. The space \mathbb{Q} contains questions about the data sets. To stress that \mathbb{Q} is not necessarily an algebra the notation $\langle x|q \rangle$ is used rather than $q(x)$ to evaluate $q \in \mathbb{Q}$ in the point $x \in \mathbb{X}$. The constant function 1 belongs to \mathbb{Q} and satisfies $\langle x|1 \rangle = 1$ for all x in \mathbb{X} .

A generalized divergence is a map $D : \mathbb{X} \times \mathbb{M} \rightarrow [0, \infty]$ satisfying the conditions

- (compatibility) for each x in \mathbb{X} is $\mu(x)$ the unique element of \mathbb{M} minimizing the divergence $m \rightarrow D(x||m)$;
- (consistency) for each m in \mathbb{M} is $0 = \inf_x \{D(x||m) : \mu(x) = m\}$.

The divergence is interpreted as the amount of information which is lost when the data set x is replaced by the model point m .

Throughout the paper we assume that there exist functions $\xi : \mathbb{M} \rightarrow \mathbb{R}$, $\zeta : \mathbb{X} \rightarrow \mathbb{R}$ and a diffeomorphism $L : \mathbb{M} \rightarrow \mathbb{Q}$ such that for all $x \in \mathbb{X}$ and $m \in \mathbb{M}$ one has

$$D(x||m) = \xi(m) - \zeta(x) - \langle x|Lm \rangle. \quad (1)$$

From the compatibility condition follows the requirement that the map $m \rightarrow \xi(m) - \langle x|Lm \rangle$ is minimal when $m = \mu(x)$. From the positivity $D(x||m) \geq 0$ and the consistency condition follows

$$\begin{aligned} \xi(m) &= \sup_x \{\zeta(x) + \langle x|Lm \rangle\} \\ &= \sup_x \{\zeta(x) + \langle x|Lm \rangle : \mu(x) = m\}. \end{aligned} \quad (2)$$

The function ζ has the meaning of an entropy function. The map L is called the logarithmic map because in the standard case (see below) it is essentially the natural logarithm. The function ξ is called the corrector [14]. We assume in what follows that it is a differentiable function.

3 Bregman divergence

The obvious example of our framework is that of a statistical model. Let \mathbb{X} be the affine space of probability distributions over a finite alphabet A . A question $q \in \mathbb{Q}$ is a real function of A . The evaluation of q in the point x is given by

$$\langle x|q \rangle = \mathbb{E}_x q = \sum_{a \in A} x(a)q(a). \quad (3)$$

Let $\theta \in \Theta \subset \mathbb{R}^n \rightarrow m_\theta \in \mathbb{X}$ be a statistical model with sufficiently nice properties so that the set

$$\mathbb{M} = \{m_\theta : \theta \in \Theta\} \subset \mathbb{X} \quad (4)$$

is a differentiable manifold.

A divergence of the Bregman type [9, 11] is defined by

$$D(x||m) = \sum_a [F(x(a)) - F(m(a)) - (x(a) - m(a))f(m(a))]$$

$$= \sum_a \int_{m(a)}^{x(a)} du [f(u) - f(m(a))], \quad (5)$$

where F is any strictly convex function defined on the interval $(0, 1]$ and $f = F'$ is its derivative. The standard case, involving the Boltzmann-Gibbs-Shannon entropy, is recovered when $F(u) = u \ln u$.

Assume that the function F is twice differentiable. The logarithmic map is given by $Lm(a) = f(m(a))$. The entropy function is $\zeta(x) = -\sum_a F(x(a))$. The consistency condition (2) follows from the convexity of the function $F(u)$. Indeed, it implies that

$$-F(x(a)) \leq -F(m(a)) - (x(a) - m(a))f(m(a)) \quad (6)$$

so that

$$\begin{aligned} \zeta(x) + \langle x|Lm \rangle &= \sum_a [-F(x(a)) + x(a)f(m(a))] \\ &\leq \sum_a [-F(m(a)) + m(a)f(m(a))] \\ &= \zeta(m) + \langle m|Lm \rangle. \end{aligned} \quad (7)$$

This implies (2). The model map μ is given by

$$\mu(x) = \arg \min_m \{\xi(m) - \langle x|Lm \rangle\}, \quad (8)$$

assuming existence and uniqueness of the minimum.

4 Generalized exponential families

Introduce now coordinates $\theta \rightarrow m_\theta$ for the model manifold \mathbb{M} . Use the notations $\xi(\theta) \equiv \xi(m_\theta)$ and $D(x||\theta) \equiv D(x||m_\theta)$. By assumption the functions $\xi(\theta)$ and $\theta \rightarrow \langle x|Lm_\theta \rangle$ are differentiable. Therefore the first derivatives

$$\frac{\partial}{\partial \theta^k} D(x||\theta) \quad (9)$$

vanish when $m_\theta = \mu(x)$.

Definition The matrix of second derivatives

$$I_{k,l}(x) = \left. \frac{\partial^2}{\partial \theta^k \partial \theta^l} D(x||\theta) \right|_{m_\theta = \mu(x)} \quad (10)$$

is the generalized Fisher information matrix.

Proposition 4.1 *The matrix $I_{k,l}(x)$ is covariant under coordinate transformations.*

Proof

Let η be a function of θ . One calculates

$$\frac{\partial^2}{\partial \theta^k \partial \theta^l} D(x||\theta) = \frac{\partial^2}{\partial \eta^m \partial \eta^n} D(x||\theta) \frac{\partial \eta^m}{\partial \theta^k} \frac{\partial \eta^n}{\partial \theta^l}$$

$$+ \left(\frac{\partial}{\partial \eta^m} D(x||\theta) \right) \frac{\partial^2 \eta^m}{\partial \theta^k \partial \theta^l}. \quad (11)$$

The latter term vanishes when $m_\theta = \mu(x)$. What remains is covariant under coordinate transformations. \square

Definition The model $\mathbb{X}, \mathbb{M}, \mathbb{Q}, D$ belongs to a generalized exponential family if the Fisher information matrix $I_{k,l}(x)$, defined by (10), is constant on the fibers $\mathcal{F}_m \equiv \{x : \mu(x) = m\}$.

The constant value is then denoted $I_{k,l}(\theta)$.

A justification of this definition follows later on from the study of the definition in the familiar context of divergencies of the Bregman type. The main advantage of the above definition is that it does not specify a particular choice of coordinates. That the Fisher information matrix is constant on the fiber \mathcal{F}_m is a scaling property. It means that locally the manifold looks always the same, independent of the point of view $x \in \mathcal{F}_m$.

5 Sufficient conditions

It is obvious to define a divergence between model points by

$$D(m||n) = \inf_x \{D(x||n) : \mu(x) = m\}. \quad (12)$$

It satisfies $D(m||n) \geq 0$. Because of the special form (1) of the divergence there follows

$$D(m||n) = \xi(n) - \sup_x \{\zeta(x) + \langle x|Ln \rangle : \mu(x) = m\}. \quad (13)$$

Using the consistency condition (2) one can write

$$\begin{aligned} D(m||n) &= \sup_x \{\zeta(x) + \langle x|Ln \rangle : \mu(x) = n\} \\ &\quad - \sup_x \{\zeta(x) + \langle x|Ln \rangle : \mu(x) = m\}. \end{aligned} \quad (14)$$

In particular, $D(m||m) = 0$ holds.

Theorem 5.1 Assume that the following Pythagorean relation^[10] holds

$$x \in \mathcal{F}_\theta \Rightarrow D(x||\theta) + D(\theta||\eta) = D(x||\eta). \quad (15)$$

Then the model belongs to the generalized exponential family.

Proof

From (15) follows

$$I_{k,l}(x) = \frac{\partial^2}{\partial \eta^k \partial \eta^l} \Big|_{\eta=\theta} D(x||\eta) = \frac{\partial^2}{\partial \eta^k \partial \eta^l} \Big|_{\eta=\theta} D(\theta||\eta). \quad (16)$$

This shows that $I_{k,l}(x)$ is constant along the fiber \mathcal{F}_θ . \square

The Pythagorean equality (15) expresses the intuition that the projection μ on the model manifold \mathbb{M} is orthogonal.

Theorem 5.2 *If the logarithmic map is of the form*

$$Lm_\theta = -\alpha(\theta) - q_0 - \eta^k(\theta)q_k \quad (17)$$

with functions α and η^k , and questions q_0, q_k in \mathbb{Q} , then the Pythagorean relation (15) is satisfied. In particular, the model belongs to a generalized exponential family.

Proof

Introduce the abbreviation $\Phi = \xi + \alpha$. From the definition of ξ follows that

$$\Phi(\theta) = \sup_x \{ \zeta(x) - \langle x|q_0 \rangle - \eta^k(\theta) \langle x|q_k \rangle \}. \quad (18)$$

Hence Φ depends on θ only via the functions η^k . In combination with

$$D(x||\theta) = \Phi(\theta) - \zeta(x) + \langle x|q_0 \rangle + \eta^k(\theta) \langle x|q_k \rangle \quad (19)$$

and the assumption that for each x there is a unique θ minimizing $D(x||\theta)$ one concludes that the map $\theta \rightarrow \eta$ is invertible. This observation is essential to conclude that $\langle x|q_k \rangle$ is constant along the fibers \mathcal{F}_θ . One has indeed for $x \in \mathcal{F}_\theta$

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta^k} D(x||\theta) \\ &= \frac{\partial \eta^m}{\partial \theta^k} \left[\frac{\partial \Phi}{\partial \eta^m} + \langle x|q_m \rangle \right] \end{aligned} \quad (20)$$

so that

$$\langle x|q_m \rangle = -\frac{\partial \Phi}{\partial \eta^m}. \quad (21)$$

Now calculate, still assuming that $x \in \mathcal{F}_\theta$, and using that $\langle x|q_k \rangle$ is constant along \mathcal{F}_θ ,

$$\begin{aligned} D(\theta||\theta') &= \inf_x \{ D(x||\theta') : x \in \mathcal{F}_\theta \} \\ &= \Phi(\theta') - \sup_x \{ \zeta(x) - \langle x|q_0 \rangle - \eta^k(\theta') \langle x|q_k \rangle : x \in \mathcal{F}_\theta \} \\ &= \Phi(\theta') - \sup_x \{ \zeta(x) - \langle x|q_0 \rangle : x \in \mathcal{F}_\theta \} - \eta^k(\theta') \langle x|q_k \rangle \\ &= \Phi(\theta') - \Phi(\theta) + (\eta^k(\theta') - \eta^k(\theta)) \langle x|q_k \rangle \\ &= D(x||\theta') - D(x||\theta). \end{aligned} \quad (22)$$

This shows the Pythagorean relation. □

6 Justification

We now return to Section 3 which deals with the Bregman divergence. In this context we give an explicit characterisation of the generalized exponential family and show that it is satisfied by the more common definition.

Taking derivatives of (5) yields

$$\frac{\partial}{\partial \theta^k} D(x||\theta) = - \sum_a [x(a) - m_\theta(a)] \frac{\partial}{\partial \theta^k} f(m_\theta(a)). \quad (23)$$

and, assuming $\mu(x) = m_\theta$,

$$\begin{aligned} I_{k,l}(x) &= \sum_a f'(m_\theta(a)) \frac{\partial m}{\partial \theta^k}(a) \frac{\partial m}{\partial \theta^l}(a) \\ &- \sum_a [x(a) - m_\theta(a)] \frac{\partial^2}{\partial \theta^k \partial \theta^l} f(m_\theta(a)). \end{aligned} \quad (24)$$

Independence of x along \mathcal{F}_θ implies

$$I_{k,l}(\theta) = \sum_a f'(m_\theta(a)) \frac{\partial m}{\partial \theta^k}(a) \frac{\partial m}{\partial \theta^l}(a) \quad (25)$$

and

$$\sum_a [x(a) - m_\theta(a)] \frac{\partial^2}{\partial \theta^k \partial \theta^l} f(m_\theta(a)) = 0 \quad \text{for all } x \in \mathcal{F}_\theta. \quad (26)$$

One concludes that the model belongs to the generalized exponential family if the set of equations (26) holds for all x satisfying $x \in \mathcal{F}_\theta$ and the normalization condition $\sum_a x(a) = 1$. With $f'(u) = 1/u$ expression (25) reduces to the standard expression for the Fisher information matrix.

The obvious solution of (26) is that there exist coordinates $\eta(\theta)$ such that

$$\frac{\partial^2}{\partial \eta^k \partial \eta^l} f(m_\theta(a)) \quad \text{does not depend on } a. \quad (27)$$

Indeed, (26) can be written as

$$\begin{aligned} 0 &= \sum_a [x(a) - m_\theta(a)] \left(\frac{\partial^2}{\partial \eta^m \partial \eta^n} f(m_\theta(a)) \right) \frac{\partial \eta^m}{\partial \theta^k} \frac{\partial \eta^n}{\partial \theta^l} \\ &+ \sum_a [x(a) - m_\theta(a)] \left(\frac{\partial}{\partial \eta^m} f(m_\theta(a)) \right) \frac{\partial^2 \eta^m}{\partial \theta^k \partial \theta^l}. \end{aligned} \quad (28)$$

Because of the *ansatz* (27) the former term vanishes. The latter vanishes because $x \in \mathcal{F}_\theta$.

The requirement (27) is equivalent with the existence of functions q_0 and q_k such that for all a and one fixed b

$$f(m_\theta(a)) - f(m_\theta(b)) = q_0(a) + \eta^k q_k(a). \quad (29)$$

One obtains

$$\langle x | f(m_\theta) \rangle = f(m_\theta(b)) + \langle x | q_0 \rangle + \eta^k \langle x | q_k \rangle. \quad (30)$$

This expression is of the form (17) (note that $f(m_\theta(a)) = Lm_\theta(a)$).

7 Discussion

We propose to replace current definitions of generalized exponential families by one formulated in terms of a generalized Fisher information — see Section 4.

The new definition can be used in a more abstract setting of information theory, one which does not necessarily rely on probability theory.

The central tool of the present paper is an asymmetric divergence $D(x||m)$ between data sets x and model points m . Divergences of this kind occur in game theory — see for instance Section 8 of [3]. They generalize the notion of a Bregman divergence [9].

The notion of a generalized exponential family is usually formulated directly in terms of the function f appearing in the generalized divergence by an expression similar to (30). We propose here to use the divergence in the first place to define a generalized Fisher information matrix. The latter is then used to define the generalized exponential families.

In [2] the function f , occurring in (30) and defining the logarithmic map L of (30), is assumed to be of the form

$$f(u) = \int_1^u dv \frac{1}{\phi(v)}, \quad (31)$$

with ϕ positive and increasing, and is called a deformed logarithm. The ϕ -deformed exponential family is then defined by an expression of the form (17). See also [7]. The special case with $\phi(v) = v^q$ is the q -deformed logarithm considered in non-extensive statistical physics [5, 15, 16]. The corresponding exponential families coincide with Amari's α -families [6, 8].

An alternative for the Bregman divergence is the U-divergence [4]. In our notations it reads

$$D_U(x||m) = \sum_a \int_{f(x(a))}^{f(m(a))} du [g(u) - x(a)], \quad (32)$$

where U is a convex increasing function, $g = U'$ and f is the inverse function of g (note that f is the deformed logarithm, g the deformed exponential function in the language of non-extensive statistical physics). The U -model is then introduced in [4] as a generalization of the exponential model and is defined by a relation of the form (17).

References

- [1] O. E. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory* (J. Wiley and Sons, New York, 1978).
- [2] J. Naudts, *Estimators, escort probabilities, and phi-exponential families in statistical physics*, J. Ineq. Pure Appl. Math. **5** (2004) 102.
- [3] P. D. Grünwald and A. P. Dawid, *Game Theory, Maximum Entropy, Minimum Discrepancy And Robust Bayesian Decision Theory*, Ann. Stat. **32** (2004) 1367–1433.
- [4] S. Eguchi, *Information geometry and statistical pattern recognition*, Sugaku Expositions (Amer. Math. Soc.) **19** (2006) 197–216 (originally Sūgaku 56 (2004) 380 in Japanese).

- [5] J. Naudts, *The q -exponential family in statistical physics*, Cent. Eur. J. Phys. **7** (2009) 405–413.
- [6] S. Amari and A. Ohara, *Geometry of q -Exponential Family of Probability Distributions*, Entropy **13** (2011) 1170–1185.
- [7] G. Pistone, *Marginal Polytope of a Deformed Exponential Family*, arXiv:1112.5123v1.
- [8] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Translations of Mathematical Monographs (Oxford University Press, Oxford, UK, 2000) (originally in Japanese (Iwanami Shoten, Tokyo, 1993)).
- [9] L.M. Bregman, *The relaxation method to find the common point of convex sets and its applications to the solution of problems in convex programming*, USSR Comp. Math. Math. Phys. **7** (1967) 200–217.
- [10] I. Csiszar, *I-Divergence Geometry of Probability Distributions and Minimization Problems*, Ann. Prob. **3** (1975) 146–158.
- [11] S. Amari and A. Cichocki, *Information geometry of divergence functions*, Bull. Pol. Acad. Sc.: Techn Sc. **58** (2010) 183–195.
- [12] J. Naudts and B. Anthonis, in preparation.
- [13] J. Naudts and B. Anthonis, *Data set models and exponential families in statistical physics and beyond*, Mod. Phys. Lett. B**26** (2012) 1250062.
- [14] F. Topsøe, *Exponential Families and MaxEnt Calculations for Entropy Measures of Statistical Physics*, arXiv:0710.1701.
- [15] C. Tsallis, *Introduction to nonextensive statistical mechanics* (Springer Verlag, 2009).
- [16] J. Naudts, *Generalised Thermostatistics* (Springer Verlag, 2011).